Modeling ‘Leave’ Vote Proportions in the 2016 UK Referendum

Section 1. Introduction and Exploratory Analysis

On 23rd June 2016, a referendum was held in the UK to decide whether to remain part of the European Union (EU), with 51.9% of voters supporting withdrawal from the EU. Martin Rosenbaum’s seminal analysis (BBC, 2017) identified correlations between Leave votes and socioeconomic factors (e.g., education levels, ethnic composition) using ward-level voting data. However, his study focused on univariate associations, instead of the joint influence of interacting variables.

In our report, we aim to use the data on the first 803 wards to build a statistical model and understand the social, economic and demographic characteristics that are associated with the voting outcome for a ward, as well as estimate the proportion of ‘Leave’ votes in each of the 267 wards for which we don’t have this information.

In the following pages, we will firstly conduct the exploratory analysis, then attempt to re-define some of variables and check the correlations of them and select the suitable variables. Finally, the LM, GLM, GAM models will be established and the estimation of the 267 wards will be made.

1.1 Distribution of ‘Leave’ proportion

To investigate the proportion of Leave in each ward, we firstly created a new variable called **LeaveProp** (= Leave / NVotes).

After calculation, the dependent variable LeaveProp (%) we are concerned about has a distribution of values between 13.16 and 78.97, with a median of 52.39 and an average of 54.82.

From the density distribution histogram (Figure1), it can be seen that the distribution of this variable is not symmetrical. From the QQ plot (Figure1), it is found that the distribution of LeaveProp deviates from normal and there is a light-tail phenomenon, especially the right thick tail, indicating that higher ‘Leave’ proportion values have a higher frequency than normal distribution.

In addition, its skewness and kurtosis values were calculated to be -0.56 and 2.56, respectively, indicating that LeaveProp is slightly skewed to the left and exhibits Platykurtic.

1.2 Covariables process

The variables given in the dataset can be roughly divided into the following groups, and there may be some logical correlations between variables in the same group.

(1) ID (Ward ID number, from 1 to 1070)

(2) Voting: NVotes (Total number of votes), Leave (Number of 'Leave' votes), Postals (whether postal votes were mixed in with the data prior to counting)

(3) Geography: RegionName, AreaType

(4) Age: MeanAge, AdultMeanAge, Age_0to4, Age_5to7, Age_8to9, Age_10to14, Age_15, Age_16to17, Age_18to19, Age_20to24, Age_25to29, Age_30to44, Age_45to59, Age_60to64, Age_65to74, Age_75to84, Age_85to89, Age_90plus

(5) Ethnicity: White, Black, Asian, Indian, Pakistani

(6) Housing situation: Owned, OwnedOutright, SocialRent, PrivateRent

(7) Education level (Low to High): NoQuals, L1Quals, L4Quals_plus

(8) Employment status: Students, Unemp, UnempRate_EA, HigherOccup, RoutineOccupOrLTU

(9) Economic indicators: Deprived, MultiDepriv

(10) Social Grades: C1C2DE, C2DE, DE

(11) Population: Residents, Households, Density (permanent residents per hectare)

In each variable groups, we choose one representative variable and check the correlations between them (Figure2). The plots show there exists collinearity among variables. Then produce some of the variables and combine similar variables to remove the possible collinearity.

(1) Age: According to *Who can vote in UK elections?*, only individuals aged 18 and over were legally eligible to vote in the referendum. Therefore, we used AdultMeanAge instead of MeanAge to analyze its relationship with LeaveProp, as it better reflects the age structure of the actual voting population.

We also grouped the population into three bands:, Young (ages 18 to 29), Working (30 to 64), Retired (65 and over), and the Children (ages 0-17) group is not considered in this study, based on typical life stages and eligibility to vote. Draw the scatter plots to show their relationships. Since the three variables are related, keep only Young since the young groups behaves different with other with working experience.

(2) Ethnicity: Since White resident is relatively largest race and Asian and Indian, Pakistani exists cross inclusion relationships, we define a new variable **Asian** (1 – White - Black) to note the Asian race. Also note that, from now on, we refer Asian as a whole including variables Asian, Indian and Pakistani, i.e. not just the given variable Asian. Then draw the scatter plots of White, Black and Asian with LeaveProp, respectively. Draw the scatter plots to show their relationships. Since the three variables are highly related, only keep White since white people has the highest proportion.

(3) Population: Since Residents and Households are highly positive related, and Density is also related with them. However, their described features are similar but not identical, and removing variables directly seems not a good chocie. Therefore, the PCA method is used and select Population.PCs1, Population.PCs2 (Figure3) to represent the original three variables.

(4) Owned or Rent: Since Owned contains OwnedOutright, remove OwnedOutright. Considering both SocialRent and PrivateRent represent renting house, create a new variable Rent = SocialRent + PrivateRent.

(5) Qualification: Considering both L1Quals and L4Quals_plus represent having qualifications and are opposite to NoQuals, combine the proportion with qualifications as Quals + L1Quals + L4Quals_plus.

(6) Occupation: Remove the UnempRate_EA since it is highly related with Unemp.

(7) Area: The RegionName contains North East, North West, Yorkshire and The Humber, East Midlands, West Midlands, East of England, London, South East and South West. It is reasonable that some regions are similar and have the similar voting preferences. Draw the boxes plot for LeaveProp with different regions (Figure4). Considering the actual geographical location and the proportion of voting to leave, combine and remap the regions: East Midlands and West Midlands are renamed as Midlands; North East, North West, Yorkshire and The Humber are renamed as Northern England; South East and South West are renamed as Southern England; East of England is renamed as Eastern England; London keeps unchanged.

According to the mapping rules, create the new variable NewRegion.

1.3 Interactions

Considering the existence of some categorical variables in the dataset, check if there are any interaction effects between variables. The Figure5 shows one example, and all of them can be seen in coding file of RegionName and Postals with other variables. The slopes of regression lines are different indicating that there may be some interaction effects between these variables, so taking them into consideration when fitting models is necessary.

In detail, LM contains the interaction terms of RegionName:AdultMeanAge, RegionName:White , RegionName:Owned, RegionName:NoQuals, RegionName:Unemp, RegionName:Deprived, Postals:Deprived, Postals:C1C2DE.

1.4 Standardization

Considering that the scales of variables vary, standardize all numerical variables within the range of 0-1 before establishing the model.

Section 2. Model Development and Compare

2.1 Linear regression model (LM)

Establish the full linear model with all variables discussed in previous text and the interaction are also added in the model according to the interaction checks.

Then the step method has been used and the AIC criteria are compared. For the simplified model, Owned, HigherOccup, RoutineOccupOrLTU, the interaction term Owned with NewRegion, Deprived with Postals.

The model is significant with adjusted R-squared of 0.884.

2.2 Generalized Linear Model (GLM)

Since the response variable is the proportion of voting to leave with the range of 0-1, the GLM method is used. The quasibinomial distribution and logit link are applied in model. The variables set is same with LM model when fitting the full model.

From the summary output, it shows that AdultMeanAge, Owned, Rent, NoQuals, HigherOccup, RoutineOccupOrLTU, C1C2DE, Postals and interaction terms of Deprived with Postals, C1C2DE with Postals are not significant. So remove these vairiables and re-fit the reduced model. The ANOVA table is calculated to compare the full model and reduced model. The F-statistic is 1.28 and p-value is 0.26, showing there is no significant differences between the two. Therefore, the simplified model with less variables are preferred.

2.3 Generalized Additive Model (GAM)

To adopt to the complex nonlinear terms of covariates, the GAM model has been established to capture the quantitative relationship between the Leave proportion and the nonlinear terms. The quasibinomial distribution is assumed in the model. Use smooth terms for all variables except interaction terms in the end to fit the full GAM.

The effective degrees of freedom (EDF) measure the complexity of each smooth term in the GAM model. A larger EDF value suggests that the fitted curve is more flexible, while a value close to one indicates a near-linear relationship. From EDF of the model summary, it can be seen that most smooth terms have high nonlinearity, while the variable "Rent" is approaching linearity. Then fit the reduced GAM model with linear "Rent", the NoQuals and RoutineOccupOrLTU are also removed since they are not significant and their EDF values are small. The ANOVA table is calculated to compare the full model and reduced model. The F-statistic is 2.99 and p-value is 0.10, showing there is no significant differences between the two. Therefore, the simplified model with less variables are preferred.

2.4 Model Comparison

In previous study, LM, GLM and GAM are established. To determine which one to use as the final model, and model diagnosis on the training dataset are considered comprehensively.

(1) Model diagnosis

In order to check model assumptions, draw the diagnosis plots for the three models (Figure 6). For LM, there are some violations on the assumption, showing non-normal distribution with heavy tails. There are also some outliers according to the plot. The diagnosis plot for GLM is similar with LM and there are also violations on model assumptions. The normal QQ plot and histogram of residuals show that the GAM's residual distribution is symmetric, but it also has heavy tails. The scatterplot residuals with linear predictor of generalized additive model demonstrates that the assumption of linearity and constant variance does hold.

(2) Evaluation indicators

To evaluate fitting effects, the three indicators: RMSE, MAE and R-squared are calculated. It is worth noting that, we cannot get the R-squared value for GLM and GAM. Therefore, the pseudo R-squared based on the deviation is calculated for GLM and GAM, while for LM, the adjusted R-squared is used directly.

The numeric comparisons are shown in Table1. It can be seen that GAM has the lowest RMSE (0.0429), MAE (0.0321), and pseudo R-squared (0.9111), indicating that GAM provides the best overall performance among the models considered. The low values of RMSE and MAE suggest that the model's predictions are highly accurate, with minimal deviation from the observed values. Additionally, the high pseudo R-squared value indicates that the GAM explains a significant proportion of the variance in the response variable, demonstrating its strong explanatory power. These results highlight the GAM's ability to effectively capture complex, non-linear relationships in the data, making it the most suitable choice for this analysis.

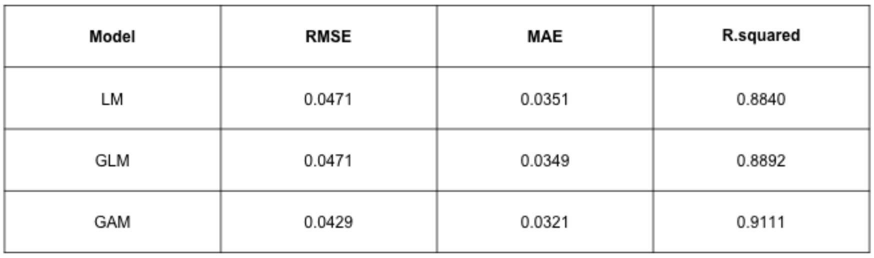
Section 3. Conclusion

In this study, the GAM model is chosen as the best model to predict the proportion of voting to leave EU. From the model we can know that nonlinear relationships and interaction terms are of great significance for predicting problems.

In conclusion, the analysis reveals that a variety of demographic, socioeconomic, and geographic factors play a significant role on voting to leave EU. Key influences include age distribution and generational differences, ethnic diversity, housing arrangements, educational attainment, and employment patterns. Additionally, economic conditions such as deprivation levels and social grade classifications contribute to the likelihood of voting to leave or remain. Regional interactions also significant in decision-making. These findings underscore the multifaceted nature of voter preferences, where individual and community-level characteristics combine to influence outcomes in voting to leave or remain.

However, several limitations should be noted. The models were trained and evaluated on the same dataset, which may result in an overestimation of predictive accuracy. Although dimension reduction and variable selection techniques were applied, some multicollinearity between predictors may remain, potentially affecting model stability. Moreover, while the GAM model captures nonlinear patterns effectively, it may risk overfitting in areas with sparse data, especially under a scoring rule that penalizes uncertainty. Future work could involve validating the models on independent datasets and refining feature selection to enhance generalizability.

Table1: Model comparison



Model	RMSE	MAE	R.squared
LM	0.0471	0.0351	0.8840
GLM	0.0471	0.0349	0.8892
GAM	0.0429	0.0321	0.9111

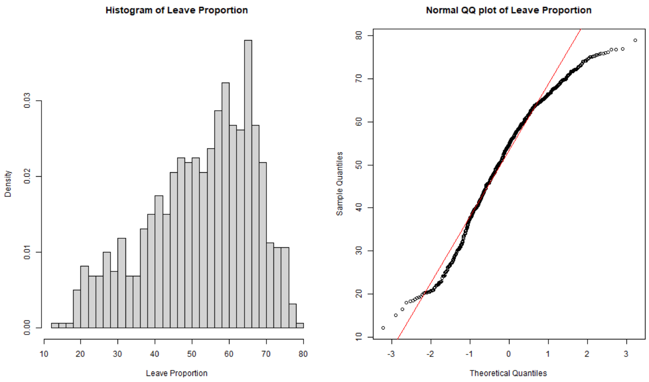


Figure1: Distribution of LeaveProp

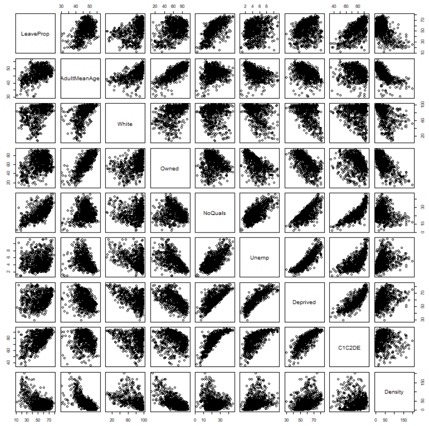


Figure2: Correlation among variables

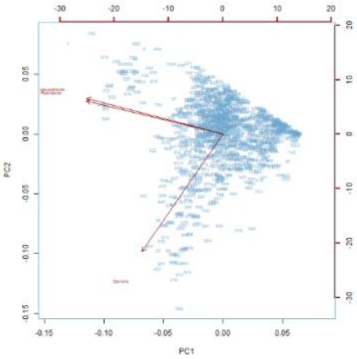


Figure3: PCA method to produce population variables

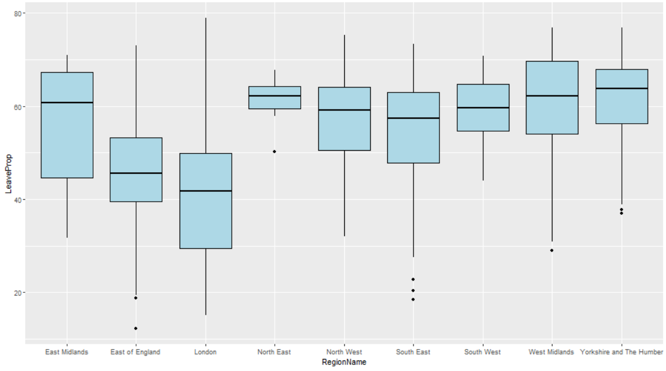


Figure4: Boxplot for LeaveProp with Different Regions

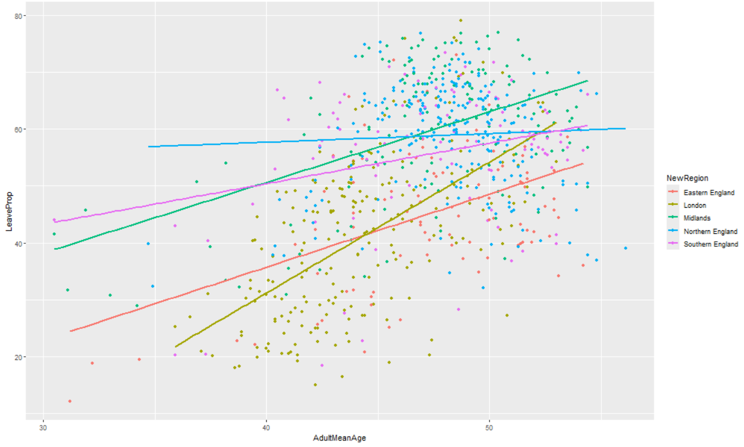


Figure5: Interaction example

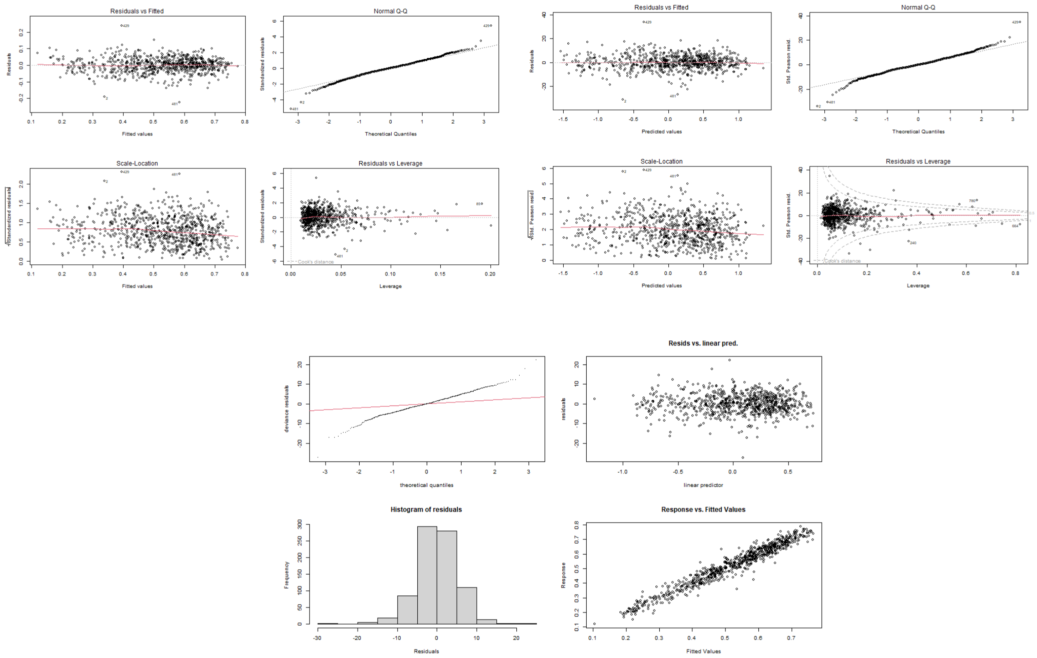


Figure6: Model diagnosis for LM, GLM and GAM

Reference

[1] BBC News: Local voting figures shed new light on EU referendum, https://www.bbc.co.uk/news/uk-politics-38762034.

Contribution Sheet

All group members contributed equally to the completion of this project.

We confirm that all members have made substantial and comparable contributions to the work submitted.

student ID

23168744

23011899

22031241

22069847